UNIVERSIDAD CARLOS III DE MADRID working papers

UC3M Working Papers
Statistics and Econometrics
15-04
ISSN 2387-0303
April 2015

Departamento de Estadística
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34) 91 624-98-48

Bayesian Linear Regression with Conditional Heteroskedasticity

Yanyun Zhao *

Abstract

In this paper we consider adaptive Bayesian semiparametric analysis of the linear regression model in the presence of conditional heteroskedasticity. The distribution of the error term on predictors are modelled by a normal distribution with covariate-dependent variance. We show that a rate-adaptive procedure for all smoothness levels of this standard deviation function is performed if the prior is properly chosen. More specifically, we derive adaptive posterior distribution rate up to a logarithm factor for the conditional standard deviation based on a transformation of hierarchical Gaussian spline prior and log-spline prior respectively.

Keywords: Bayesian linear regression, conditional heteroskedasticity, rate of convergence, posterior distribution, adaptation, hierarchical Gaussian spline prior, log-spline prior.

*Department of Statistics, Universidad Carlos III de Madrid. Email: yanyun.zhao@uc3m.es.

Bayesian Linear Regression with Conditional Heteroskedasticity

Yanyun Zhao
Departamento de Estadística, Universidad Carlos III de Madrid

Abstract

In this paper we consider adaptive Bayesian semiparametric analysis of the linear regression model in the presence of conditional heteroskedasticity. The distribution of the error term on predictors are modelled by a normal distribution with covariate-dependent variance. We show that a rate-adaptive procedure for all smoothness levels of this standard deviation function is performed if the prior is properly chosen.More specifically, we derive adaptive posterior distribution rate up to a logarithm factor for the conditional standard deviation based on a transformation of hierarchical Gaussian spline prior and log-spline prior respectively.

Keywords: Bayesian linear regression, conditional heteroskedasticity, rate of convergence, posterior distribution, adaptation, hierarchical Gaussian spline prior, log-spline prior.

1 Introduction

We consider Bayesian estimation of the linear regression model that imposes conditional moment restrictions. A useful framework like $E(Y|X) = X'\beta_0$ or $Y = X'\beta_0 + \varepsilon$, $E(\varepsilon|X) = 0$ is widely formulated to analyze a number of statistical and econometric models. It is well-known that the procedure of estimating the parameters of interest could be expected to be efficient provided more information about the conditional error distribution is known. In this paper we propose a Bayesian semiparametric method for consistent estimation of the regression coefficients and the conditional standard deviation when the error term is subject to a normal distribution with associated variance that is dependent on covariates.

The primary purpose of this paper is to investigate the asymptotic frequentist properties of the corresponding posterior distribution by putting a prior on the regression coefficients and the standard deviation in this linear model. An analysis of the asymptotic behavior of Bayesian methods in the infinite-dimensional statistical models is important, such as posterior consistency, rate of posterior convergent, rate-optimality and adaptation properties and Bernstein-von Mises phenomenons, which reflect a sense of Bayesian robustness, namely the prior does not have an impact on the posterior distribution too much when the amount of information collected in the data or the number of observations grows indefinitely.

In recent years, there has been substantial research in Bayesian nonparametrics on the development of these mathematical, asymptotical theory for a wide range of statistical models, see, for example, Ghosal et al. (1999, 2000); Ghosal and van der Vaart (2001, 2007b,a), to name a few. However, it has been studied very little in the linear models with predictor dependent conditional variance of the error terms. Norets (2015) established a semiparametric version of Bernstein-von Mises theorem under misspecification: the posterior credible regions of the regression coefficients are asymptotically equivalent to the frequentist ones and also this posterior inference is efficient even though the data generating process is not normal. Pelenis (2014) considered the kernel stick-breaking mixtures to model the conditional error distribution and demonstrated posterior consistency of the conditional error density and the finite regression coefficients for these kernel mixture priors. Also, Wang (2013) studied posterior consistency for the heteroscedastic nonparametric regression models by relaxing the assumptions of linearity in the model, with a substitution of an unknown, smooth regression function. There is a noticeable absence of rate adaptation results in these regression setting.

In the present paper, we plug this gap and take up the investigation of this rate adaptive procedure, in order to provide a theoretical underpinning of the Bayesian inference approach to explore the possible accuracy at maximum capacity and assess the well-balanced spread of the underlying prior distribution across a continuum of regularities of the functions considered. Adaptive convergence rates for Bayesian nonparametric estimation in various statistical models have been established by Huang (2004), Scricciolo (2006), Belitser and Ghosal (2003), van der Vaart and van Zanten (2009), Rousseau (2010), Kruijer et al. (2010), de Jonge and van Zanten (2010, 2012), Shen and Ghosal (2012), Shen et al. (2013), Norets and Pati (2014) and Belitser and Serra (2014), among others.

A broad class of priors have been explored to yield adaptation across all smoothness levels. Recently, priors based on splines have received much attention for the construction of probability distribution on the infinite-dimensional spaces. Various groups of researchers have worked with univariate splines or its corresponding tensor-product splines in the multivariate case as a useful block to construct a prior. For example, Huang (2004) built a prior on the discrete mixture of splines to develop a theorem on adaptive convergence rates in the context of regression and density estimation. de Jonge and van Zanten (2012) discussed priors on multivariate functions by choosing an appropriate probability distribution on the partition size and Gaussian prior on B-spline coefficients in the tensor-product B-spline expansions. Shen and Ghosal (2012) constructed a prior using finite random splines with a prior distribution on the number of terms. Belitser and Serra (2014) investigated an extension of these results involving spline based priors by endowing a probability distribution on the location of the knots instead of assuming them to be equally spaced. This enables us to build a wide spectrum of priors on the conditional standard deviation of the regression error terms. It is widely known that the posterior distribution contracts at a rate of the order $n^{-\alpha/(2\alpha+d)}$ (up to an additional logarithm factor) for a α-smooth functions of d-variables, which agrees with the optimal rate of the estimators in the frequentist context. In other words, a fully rate-adaptive

procedure can be obtained across all smoothness levels if that holds. One possible explanation about this phenomenon is that there is a sufficiently large amount of prior mass around the function of interest with total smoothness levels. We will show that the corresponding posterior converges at the optimal rate up to a logarithm factor without the priori knowledge of the smoothness levels of the conditional standard deviation.

From the practical point of view, diverse algorithms for normal linear regression with predictor dependent variance have been exhibited in Yau and Kohn (2003) and Chib and Greenberg (2013) which considered transformed splines to model the variance and Goldberg et al. (1997) where a transformed Gaussian process prior was considered. Markov Chain Monte Carlo simulations carried out in these papers performs well in these flexible covariance dependent cases. Here we center on the theoretical aspects in Bayesian normal regression models.

The paper is organized as follows. In Section 2 we give a general overview of the notation and a brief account of the model settings. In Section 3 we provide a preliminary review on the notions of spline functions, univariate B-splines and tensor-product B-splines as well as its associated approximation properties. In Section 4, we show that the optimal posterior convergence rate can be achieved using two types of spline priors: one based on conditional Gaussian tensor-product spline prior or a hierarchical Gaussian spline prior and the other built on log-spline prior that stems from finite random spline expansion with a random number of terms. We conclude with a brief discussion and some technical lemmas, all containing proofs as well as auxiliary theorems are delegated to the Appendix.

2 General model setup

In this Section, we take a detailed description of the notation and then describe our model.

2.1 Notation

For any $a \in \mathbb{R}$, denote $\lfloor a \rfloor$ to be the largest integer strictly smaller than a. Similarly, define $\lceil a \rceil$ to be the smallest integer which is strictly greater than a.

Let $\eta = (\beta, \sigma)$ and the true value $\eta_0 = (\beta_0, \sigma_0)$. Denote the conditional density function $N(\beta, \sigma^2(x))$ by $f_{x\eta}$ and let $f_{x\eta_0}$ be the true conditional density function $N(\beta_0, \sigma_0^2(x))$. The Kullback-Leibler divergence between η and η_0 is then defined as,

$$K(\eta, \eta_0) = \int_{\mathcal{X}} \int_{\mathcal{Y}} f_{x\eta_0}(y) \log \frac{f_{x\eta_0}(y)}{f_{x\eta}(y)} \, dy \, dG_0(x), \tag{1}$$

$$V(\eta, \eta_0) = \int_{\mathcal{X}} \int_{\mathcal{Y}} f_{x\eta_0}(y) \left(\log \frac{f_{x\eta_0}(y)}{f_{x\eta}(y)} \right)^2 dy \, dG_0(x), \tag{2}$$

where $\mathcal{X}, \mathcal{Y}$ are the domains that will be specified later and $G_0(\cdot)$ is a general distribution function. The ε-Kullback-Leibler neighborhood around η_0 is expressed as,

$$K_\varepsilon(\eta_0) = \{\eta : K(\eta, \eta_0) < \varepsilon\}. \tag{3}$$

We define the Hellinger metric between η and η_0 as,

$$d_H(\eta, \eta_0) = \int_{\mathcal{X}} \int_{\mathcal{Y}} \left(\sqrt{f_{x\eta}(y)} - \sqrt{f_{x\eta_0}(y)} \right)^2 dy \, dG_0(x). \tag{4}$$

We use the natural L^2-norm with respect to the distribution function $G_0(\cdot)$ to measure the distance between η and η_0:

$$d_2(\eta, \eta_0) = \left\{ \int_a^b \left([(\beta - \beta_0)^T x]^2 + [\sigma(x) - \sigma_0(x)]^2 \right) dG_0(x) \right\}^{1/2}, \tag{5}$$

and denote the neighborhood of η_0 with respect to the distance function $d_2(\eta, \eta_0)$ as follows:

$$U_\varepsilon(\eta_0) = \left\{ (F, \sigma) : \int_a^b \left([(\beta - \beta_0)^T x]^2 + [\sigma(x) - \sigma_0(x)]^2 \right) dG_0(x) > \varepsilon \right\}. \tag{6}$$

We use the notation $\lesssim$ to stand for somewhat inequality up to a constant. To compare two function, for example, g_1, g_2, we denote $g_1 \lesssim g_2 \lesssim g_1$ by $g_1 \asymp g_2$. The covering number of a set $\mathscr{P}$ equipped with some metric d, denoted by $N(\varepsilon, \mathscr{P}, d)$, is viewed as the minimum number of d-balls with radius ε needed to cover the set $\mathscr{P}$. The metric entropy number of the set $\mathscr{P}$, denoted by $\log N(\varepsilon, \mathscr{P}, d)$, is defined as the logarithm of its associated covering number. Let $\|\cdot\|_2$ and $\|\cdot\|_\infty$ denote the Euclidean norm and supremum norm respectively.

We now take a brief account of the definitions in the context of multivariate functions, especially describe the appropriate notions of smoothness in this multivariate case. Let's denote the space of continuous functions f on $[0,1]^d$ by $C\left([0,1]^d\right)$, equipped with the supremum norm $\|f\|_\infty$. For a multi-index $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_d)$, let the sum $|\alpha| = \sum_{i=1}^d \alpha_i$ and the mixed partial derivative operator is defined as,

$$D^\alpha = \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}. \tag{7}$$

For $\alpha > 0$, the Hölder space $C^\alpha\left([0,1]^d\right)$ stands for the collection of functions f on $[0,1]^d$ with mixed partial derivative $D^r f \in C\left([0,1]^d\right)$ of all orders up to $|r| \leq \lfloor \alpha \rfloor$ satisfying,

$$|D^r f(x) - D^r f(y)| \leq C \|x - y\|_2^{\alpha - \lfloor r \rfloor}, \tag{8}$$

for some positive constant C, each $x, y \in [0,1]^d$. Meanwhile, denote the norm on the Hölder class $C^\alpha\left([0,1]^d\right)$ by,

$$\|f\|_{C^\alpha([0,1]^d)} = \|f\|_\infty + \sum_{r:\, |r| = \lfloor \alpha \rfloor} \|D^r f\|_\infty. \tag{9}$$

2.2 Restricted moment models

Suppose we observe a real-valued sample $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ where X_i is a d-dimensional covariate, Y_i is the response variable and $(X_i, Y_i) \sim P_0$ for $i = 1, 2, \ldots, N$. The data generating process satisfies $Y|X = x \sim N(x'\beta_0, \sigma_0^2(x))$ for some unknown true parameter

$\beta_0 \in \Theta \subset \mathbb{R}^d$ and unknown true conditional variance function $\sigma_0^2(\cdot) : [0,1]^d \to (0,\infty)$ and all $x \in \mathscr{X} = [0,1]^d$. In other words, this linear model could be described as,

$$Y_i = X_i'\beta_0 + \varepsilon_i, \quad i = 1, 2, \ldots, n. \tag{10}$$

where error variables $\varepsilon_i | X_i = x_i \sim N(0,\, \sigma_0^2(x_i))$ for all $x_i \in [0,1]$, $i = 1, 2, \ldots, n$. In this semiparametric model, the unknown parameters are $(\beta,\, \sigma(\cdot))$ where the finite-dimensional parameter β is of interest and $\sigma(\cdot)$ is the infinite-dimensional nuisance parameter. Our model could be rewritten as $(\Theta \times \mathcal{M},\, \mathscr{B} \times \mathscr{F})$ equipped with Borel σ-algebras $\mathscr{B}$ and $\mathscr{F}$ on Θ and $\mathcal{M}$ respectively, where,

$$\mathcal{M} = \{\sigma(\cdot) :\, [0,1]^d \to (\underline{\sigma}, \overline{\sigma})\}. \tag{11}$$

is a polish space on $\mathscr{X}$ and also is assumed to contain the true conditional standard deviation σ_0. Let Π denote the total prior for the pair $(\beta,\, \sigma)$ on $(\Theta,\, \mathcal{M})$ which is defined by $\Pi(d\beta, d\sigma) = \Pi_\beta(d\beta) \times \Pi_\sigma(d\sigma)$ where Π_β and Π_σ are corresponding independent priors on β and σ respectively. Here we leave the distribution of covariates denoted by $G_0(\cdot)$ unspecified since it is ancillary and also of our interest is to focus on the conditional distribution. The corresponding posterior distribution for (β, σ) given the data $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ is denoted by,

$$\Pi(\cdot | (X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)).$$

In view of Bayes' theorem, the posterior is given by the expression,

$$\Pi(B | (X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)) = \frac{\int_B L(\beta, \sigma; (X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n))\, \Pi(d\beta, d\sigma)}{\int L(\beta, \sigma; (X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n))\, \Pi(d\beta, d\sigma)}, \tag{12}$$

where the likelihood function $L(\beta, \sigma; (X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n))$ could be written as,

$$\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma(X_i)} \exp\left(-\frac{(Y_i - X_i'\beta)^2}{2\sigma^2(X_i)}\right). \tag{13}$$

Usually the posterior mean can be regarded as a Bayesian estimator of the unknown pair (β_0, σ_0). If this Bayesian estimator is consistent, the further concern is then of interest to consider the finer aspects of this posterior distribution or quantify the rate at the which it contracts around the true unknown parameter, namely, posterior convergence rate. More precisely, for a given positive sequence (ε_n) going to zero, the posterior distribution is said to converge to the Dirac-mass at (β_0, σ_0) at the rate ε_n, if, as $n \to \infty$,

$$\Pi\big\{(\beta, \sigma) :\, d_H((\beta, \sigma), (\beta_0, \sigma_0)) > M\varepsilon_n \;\big|\; (X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\big\} \longrightarrow 0 \text{ in } P_0^n\text{-probability}, \tag{14}$$

for a sufficiently large $M > 0$. Here this assertion of the definition is in-probability statement that holds under the true distribution P_0 governed by the true parameter pair (β_0, σ_0).

The main objective is to construct some priors for $\Theta \times \mathcal{M}$ to show the corresponding posterior converges at an optimal rate at $(\beta_0, \sigma_0(\cdot)) \in \Theta \times \mathcal{M}$. Here the prior does not depend on the information about the unknown smoothness levels of the true conditional standard deviation function $\sigma_0(\cdot)$. So the so-called rate-adaptive procedure is obtained across all the regularity levels.

3 A preliminary introduction to Splines

In this Section, we will provide a general overview on spline function supported on hyper cube following by a brief introduction on the splines defined on the unit interval $[0, 1]$. More extensive treatment on this subject could be found in Schumaker (2007).

3.1 Spline function on unit interval

A spline function on $[0, 1]$ is essentially viewed as an generalization of the polynomial function on the unit interval. It is a piece polynomial function but enjoy the properties of global smoothness on its domain.

More specifically, let q, K be two fixed natural numbers and partition the unit interval $[0, 1]$ into K equally spaced subintervals $[(k-1)/K, k/K]$ for $k = 1, 2, \ldots, K$. Consider a spline function with the order q greater than 2, that is, all polynomials with its domain coinciding with one of those subintervals are of the degree smaller than $q - 1$ and this spline function is globally $q - 2$ times continuously differentiable on $[0, 1]$.

Let S_K be the collection of all splines of order q with simple knots at the points $\{k/K : k = 1, \ldots, K-1\}$. It can be seen that S_K forms a $J = (q + K - 1)$-dimensional liner space. The so-called B-splines $B_1^K, B_2^K, \ldots, B_J^K$, which can be found in de Boor (2001), are used to give a convenient basis in this space. The concrete function forms of these B-splines are negligible to us. The primary properties of these B-splines closely used in this paper are that B-splines are always nonnegative, each basis function is supported on a tiny interval with its length at most q/K and the sum of all B-splines evaluated at any given point in the domain is equal to one. In other words, they constitute a partition of unity, i.e.

$$\sum_{i=1}^{J} B_i^K(x) = 1,$$

for each $x \in [0, 1]$.

3.2 Tensor-product spline on $[0, 1]^d$

In this Subsection we introduce spline functions on multi-dimensional domains with the help of multivariate polynomials. The construction of the linear space of such multivariate splines relies heavily on the spline space S_K in the unit interval described above. In fact, this linear space on $[0, 1]^d$ is a tensor-product of those univariate linear space on $[0, 1]$. More precisely, a unique direction denoted by a variable is assigned to each linear space in the tensor-product and then we obtain the multivariate polynomials supported on some tiny rectangles by taking the multiplication of polynomials with respect to one single variable defined on some small intervals.

Accordingly, the convenient basis for the linear space of tensor-product splines is the tensor-product B-splines, which equal to the products of the corresponding B-splines on $[0,1]$. Hence the tensor product space has dimension $(q+K-1)^d$, for example, in the construction of the space S_K defined above. The advantage of introducing tensor-product B-splines is that they inherit the nice properties that univariate B-splines have as we shall see below.

In what follows, we consider a d-fold tensor-product space $\mathcal{S}_K = S_K \otimes \cdots \otimes S_K (d \text{ times})$ of tensor-product splines defined on the unit cube $[0,1]^d$, that is partitioned equally into m^d cubes $I_{k_1} \times \cdots \times I_{k_d}$. A function $s : [0,1]^d \to \mathbb{R}$ is defined to be a tensor-product spline in $\mathcal{S}_K$ if for each such tiny cube, s possesses the following multivariate polynomial form,

$$\sum_{k_1=0}^{q-1} \cdots \sum_{k_d=0}^{q-1} c_{k_1 \ldots k_d} \, x_1^{k_1} \cdots x_d^{k_d}. \tag{15}$$

As was the case in the univariate spline space, the basis in $\mathcal{S}_K$ is provided by the so-called tensor-product B-splines as follows,

$$B^K_{j_1 \ldots j_d}(x_1, \ldots, x_d) = B^K_{j_1}(x_1) B^K_{j_2}(x_2) \cdots B^K_{j_d}(x_d). \tag{16}$$

It can be shown that $\mathcal{S}_K$ has dimension $(q+K-1)^d$ and these multivariate B-splines also form a partition of unity,

$$\sum_{j_1=1}^{J} \cdots \sum_{j_d=1}^{J} B^K_{j_1 \ldots j_d}(x_1, \ldots, x_d) = 1, \tag{17}$$

for all $x_i \in [0,1]$, $i = 1, 2, \ldots, d$.

3.3 Approximation properties of tensor-product B-splines

It is well-known that the univariate B-splines in the space S_K could approximate any function of interest in $C^\alpha[0,1]$, for example, at the rate $J^{-\alpha}$ where $J = q+K-1$. In other words, any function with a smoothness level α in $C^\alpha[0,1]$ could be approximated by a couple of B-splines, $B^K_1, B^K_2, \ldots, B^K_J$ with its associated approximation error controlled by the order $J^{-\alpha}$.

This idea also works in the multivariate case. How well tensor-product B-splines approximate the generic function is uniquely determined by the target function's smoothness level α and the dimension of the linear space $\mathcal{S}_K$ induced by the tensor-product B-splines if the order q of the splines is chosen to be larger than the smoothness level α. The approximation ability in terms of tensor-product B-splines is stated in the following lemma which provides an upper bound of the approximation error with respective to the uniform distance.

LEMMA 3.1 (Shen and Ghosal (2014)) *Let q, d, $K \in \mathbb{N}$, $\alpha \in \mathbb{R}$, $\alpha \leq q$, $J = q+K-1$. For any function $f \in C^\alpha\left([0,1]^d\right)$, there exists $\boldsymbol{\theta} = (\theta_{00\ldots0}, \ldots, \theta_{JJ\ldots J}) \in \mathbb{R}^{J^d}$ and a positive constant C_1 that only depends on q, d and α such that,*

$$\left\| f - \sum_{j_1=1}^{J} \cdots \sum_{j_d=1}^{J} \theta_{j_1 \ldots j_d} \, B^K_{j_1 \ldots j_d}(x_1, \ldots, x_d) \right\|_\infty \leq C_1 J^{-\alpha} \|D^\alpha f\|_\infty. \tag{18}$$

Furthermore, if $f > 0$, then each element of $\boldsymbol{\theta}$ could be chosen to be positive for a sufficiently large J.

4 Adaptive posterior contraction results

Splines possess excellent approximation capabilities for smooth functions in the previous Section, where the approximation error is completely controlled by the dimension of the spline space and the smoothness level. More precisely, the error becomes smaller if the dimension grows and the objective function is smoother. From the frequentist view of point, Stone (1994) showed that the maximum likelihood estimator of the function in $C^{\alpha}([0,1]^d)$ achieves the rate of convergence $n^{-\alpha/(2\alpha+d)}$. As indicated in de Jonge and van Zanten (2012), a Bayesian estimator for probability densities or the regression functions in multivariate domains under more weaker conditions also attain the optimal contraction rate $n^{-\alpha/(2\alpha+d)}$. Simultaneously, they established that a type of Gaussian process prior yields the near-optimal adaptive posterior convergence rate, up to an additional logarithmic factor when α is unknown.

In the next two Subsections, we consider spline-based priors for $\sigma(\cdot)$ in a variety of means. In Subsection 4.1, we build a hierarchical Gaussian spline prior by putting Gaussian prior weights on the coefficient and adding another hierarchical layer for the partition size involved in the tensor-product B-splines. It follows that this hierarchical procedure achieves a near-optimal adaptive contraction rate. Alternative log-spline priors with finite random tensor-product splines and a random number of terms that also achieve optimal adaptive rate of convergence will be demonstrated in Subsection 4.2.

Throughout this Section, we consider the following condition on Π_{β} :

(A1) Its support is $[\underline{\beta}, \overline{\beta}]$. For all $\varepsilon > 0$, there exists $m_1 > 0$ such that,

$$\Pi(\|\beta - \beta_0\|_2 \leq \varepsilon) \geq \exp(-m_1 d \log(1/\varepsilon)). \tag{19}$$

In fact, this is a mild assumption on the prior of β. And several ordinary distribution examples satisfy (19). More detailed and similar examples could be found in the discussion of the prior for weights vector $\boldsymbol{\theta}$ in Subsection 4.2.

4.1 Hierarchical Gaussian spline prior

In this Subsection, a class of Gaussian process, whose sample path is defined by tensor-product splines extensively discussed in the preceding Section, will be used for the construction of priors on the conditional standard deviation in our linear model.

Let $Z_{00\ldots0}, \ldots, Z_{JJ\ldots J}$ be a series of i.i.d standard normal random variables, the random pro-

cess W^K on $[0,1]^d$ was given by

$$W^K(x_1,\ldots,x_d) = \sum_{j_1=1}^{J} \cdots \sum_{j_d=1}^{J} Z_{j_1\ldots j_d} B^K_{j_1\ldots j_d}(x_1,\ldots,x_d), \quad x_i \in [0,1],\, i = 1,2,\ldots,d. \tag{20}$$

where $\{B^K_{j_1\ldots j_d}(x_1,\ldots,x_d) : j_i = 1,\ldots,J,\, i = 1,2,\ldots,d\}$ is a group of tensor-product B-spline basis of $\mathcal{S}_K$, $J = q+K-1$, K is the partition size of the knots. de Jonge and van Zanten (2012) has shown that $\{B^K_{j_1\ldots j_d}(x_1,\ldots,x_d) : j_i = 1,\ldots,J,\, i = 1,2,\ldots,d\}$ form an orthonormal basis of the reproducing kernel Hilbert space (RKHS) $\mathbb{H}^K$ associated with this Gaussian process W^K and also extensively exhibited the properties of the concentration function, which plays a crucial role in determining the posterior convergence rate regarding to this Gaussian process prior induced by the stochastic process W^K.

In order that the corresponding posterior could be guaranteed to take on the asymptotic properties, posterior consistency for example, the prior should have large enough support. The tuning parameter K then should be required to vary with the sample size as well as the regularity of the function of interest and the number of observations should also go to infinity. This prior, the law of the Gaussian spline prior W^K, depends explicitly on the unknown smoothness level of the object. So this is not desired rate-adaptive procedure.

We could remedy this problem if this partition size K is viewed as the so-called hyper parameter and itself is endowed with a separate prior. In other words, we assign a probability distribution on such an unknown tuning parameter and let the partition size be carefully selected through its posterior distribution. In the Bayesian perspective, it is natural to treat this parameter as one type of hyper parameter and let it estimated from the data via its posterior mean.

Let $\tilde{K}$ be an independent $\mathbb{N}$-valued random variable, the hierarchical Gaussian process prior is denoted by $W^{\tilde{K}}$, where $W^{\tilde{K}}|_{\tilde{K}=K}$ is described in (20). As prior on the standard deviation, we employ the law Π_σ of the process $\tilde{\Psi}(W^{\tilde{K}})$, that is a transformation of the stochastic process $W^{\tilde{K}}$, where the link function $\tilde{\Psi} : \mathbb{R} \to (\underline{\sigma}, \overline{\sigma})$ is given by,

$$\tilde{\Psi}(W^{\tilde{K}}) = \Psi(W^{\tilde{K}})(\overline{\sigma} - \underline{\sigma}) + \underline{\sigma}, \tag{21}$$

for the logistic or normal function distribution Ψ.

The following theorem follows from Theorem 4.2 in de Jonge and van Zanten (2012) that presents the general rate of contraction results for Bayesian multivariate function estimation.

THEOREM 4.1 *Assume that $w_0 = \tilde{\Psi}^{-1}(\sigma_0) \in C^\alpha([0,1]^d)$ for some integer α less than q. Let the prior Π_σ be induced by the law of the stochastic process $\tilde{\Psi}(W^{\tilde{K}})$, where the probability mass of this hyper parameter $\tilde{K}$ for each $K \geq 1$ satisfies:*

$$C_1 \exp(-D_1 K^d \log^t K) \leq P(\tilde{K} = K) \leq C_2 \exp(-D_2 K^d \log^t K), \tag{22}$$

for some constants $C_1, C_2, D_1, D_2, t \geq 0$. Suppose that for any $\varepsilon > 0$, $\log\left\{\left[\frac{\overline{\beta}-\underline{\beta}}{2\varepsilon}\right] + 1\right\} \leq n\varepsilon^2$ and also the prior for the regression coefficient Π_β satisfies $(A1)$. Let the maximal eigenvalue

of $E(X_iX_i')$ denoted by $\lambda_{max}(E(X_iX_i'))$ be bounded for $i = 1, 2, \ldots, n$. Then, for a sufficiently large constant $M > 0$,

$$\Pi_n\{\eta : d_H(\eta, \eta_0) > M\varepsilon_n | (X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\} \longrightarrow 0 \text{ in } P_0^n\text{-probability},$$

where,

$$\varepsilon_n = c(n/\log^{1\vee t} n)^{-\frac{\alpha}{d+2\alpha}} \vee n^{-\frac{\alpha}{d+2\alpha}} (\log n)^{\frac{(1\vee t)\alpha}{d+2\alpha} + (\frac{1-t}{2})+},$$

for a large enough positive constant c.

Note that if $\tilde{K}^d$ follows a Geometric distribution with $t = 0$, then condition (22) is satisfied. Here the stochastic process prior $\tilde{\Psi}(W^{\tilde{K}})$ implies a posterior rate of concentration on the space of the standard deviation functions provided the true standard deviation has regularity level α less than q. As indicated in de Jonge and van Zanten (2012), we keep the order q involved in the splines fixed so that the prior could become simpler as well as easier for simulations computationally. A common choice for q is 4 in practice.

The prior does not depend on the smoothness level α so our procedure is adaptive. If t is chosen to be equivalent to one, the the rate ε_n becomes $(n/\log n)^{-\frac{\alpha}{d+2\alpha}}$, which coincides with the optimal posterior convergence rate, up to an additional logarithm item, since the rate $n^{-\frac{\alpha}{d+2\alpha}}$ for each $\alpha > 0$ is the minimax convergence rate in the function class $C^\alpha([0,1]^d)$.

4.2 Log-spline prior

We consider a prior, in this Subsection, induced by a random series expansion in terms of tensor-product B-splines as follows:

$$W^{J,\boldsymbol{\theta}}(x) = \sum_{j_1=1}^{J} \cdots \sum_{j_d=1}^{J} \theta_{j_1 \ldots j_d} B^K_{j_1 \ldots j_d}(x_1, \ldots, x_d), \tag{23}$$

where $\boldsymbol{\theta} = (\theta_{00\ldots0}, \ldots, \theta_{JJ\ldots J})$ is a J^d-dimensional vector. A prior on h could be obtained by assigning a probability distribution on the number of items J and the associated coefficient vector $\boldsymbol{\theta}$ of tensor-product B-splines discussed in Shen and Ghosal (2012) as follows:

(A2) We consider a prior for J satisfying,

$$\exp(-c_1 j \log^{t_1} j) \leq \Pi(J = j) \leq \exp(-c_2 j \log^{t_2} j), \quad j = 1, 2, \ldots, \tag{24}$$

for some positive constants c_1, c_2 and $0 \leq t_1 \leq t_2 \leq 1$.

(A3) Given J, the prior for J^d-dimensional vector $\boldsymbol{\theta}$ satisfies for each $\|\boldsymbol{\theta}_0\|_\infty \leq H$ and a sufficiently small $\varepsilon > 0$,

$$\Pi(\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2 \leq \varepsilon) \geq \exp(-c_3 J^d \log(1/\varepsilon)), \tag{25}$$

$$\Pi(\boldsymbol{\theta} \notin [-M, M]^{J^d}) \leq J^d \exp(-c_4 M^{t_3}), \tag{26}$$

for some positive constants c_3, c_4, t_3 and sufficiently large $M > 0$.

Note that (A2) holds for Geometric, Poisson and Negative distributions when t_1, t_2 are carefully chosen. And (A3) is fulfilled if we put independent Gamma and Exponential distributions on each element of the vector $\boldsymbol{\theta}$. If the support of $\boldsymbol{\theta}$ is a bounded and closed set, then multivariate Normal and Dirichlet distributions also meet (A3). We take the law of the following stochastic process as the prior on the standard deviation σ:

$$\tilde{\Phi}(W^{J,\boldsymbol{\theta}}(x)) = \frac{e^{W^{J,\boldsymbol{\theta}}(x)}}{\int_0^1 e^{W^{J,\boldsymbol{\theta}}(x)}\, dx}(\overline{\sigma} - \underline{\sigma}) + \underline{\sigma}, \tag{27}$$

where $W^{J,\boldsymbol{\theta}}(x)$ is defined in (23). The law of the process $\tilde{\Phi}$ gives the so-called log-spline prior for the infinite-dimensional parameter σ.

We now present the result about the contraction rate of the posterior based on the product prior defined by Π_β and this log-spline prior.

THEOREM 4.2 *Let $w_0 = \tilde{\Phi}^{-1}(\sigma_0) \in C^\alpha([0,1]^d)$ and the prior for the regression coefficient β, the number of items J and the associated coefficients $\boldsymbol{\theta}$ satisfy (A1), (A2) and (A3) respectively. Suppose that the maximal eigenvalue of $E(X_i X_i')$ is bounded for $i = 1, 2, \ldots, n$. Assume that we endow a prior on σ by the law of the process $\tilde{\Phi}(W^{J,\boldsymbol{\theta}})$, then the corresponding posterior of $\eta = (\sigma, \beta)$ contracts at the rate,*

$$\varepsilon_n = n^{-\alpha/(2\alpha+d)}(\log n)^{\alpha/(2\alpha+d)-(t_2-1)/2}, \tag{28}$$

in terms of the Hellinger distance d_H.

In fact, we applies Theorem 2 in Shen and Ghosal (2012) to our linear model in the presence of the heteroscedasticity with this prior Π_η to get this result. The optimal posterior convergence rate relative to the Hellinger distance could be obtained by carefully selecting some sequences $\bar{J}_n$, J_n, M_n, $\bar{\varepsilon}_n$ that satisfy the conditions stated in Theorem 2 of Shen and Ghosal (2012) in order to balance bias and model complexity in our semiparametric model.

5 Conclusions

To summarise, we obtain an adaptive procedure in a flexible linear model with heteroscedastic normally distributed error in the presence of a conditional moment condition. More specifically, under mild restrictions on the model and priors, the posteriors of the conditional standard deviation and of the finite regression coefficients adapt to the smoothness of the underlying standard deviation function, which is assumed to be contained in a nonparametric model. This result indicated that we could implement this Bayesian procedure as if the regularity of the underlying function were known.

The alternative asymptotic property concerning in our normal linear regression model, the Bernstein-von Mises theorem, has been developed in Norets (2015). Further research is warranted for the investigation of the existence of a Bernstein-von Mises phenomenon in this

semiparametric model where the parameter of interest is the finite-dimensional regression coefficients, by directly assigning a prior on the conditional error distribution with a zero mean restriction. The estimation of the coefficients of interest in this setting that avoid the potential model misspecifications would be efficient. Particularly challenging is how to model this conditional error density with the imposition of moment restriction. Moreover, the problem is compounded by the fact that the appropriate constructions of the priors put on these conditional error densities, making it difficult to obtain the semiparametric efficiency bound.

It would be interesting to extend the adaptive concentration rate of posterior and Bernstein-von Mises theorem in our model to that in the weakly dependent data. In infinite-dimensional models, there are few results concerning these two important asymptotic properties in the weakly dependent cases. Maybe we could establish this asymptotic results under appropriate conditions on the prior, an interesting future direction.

A Useful lemmas

To prove the main theorems in Section 4, we need the following supplementary lemmas. For brevity of notations, we use the generic positive constant C throughout this Appendix.

LEMMA A.1 *If $x > 0$, then the following inequality holds.*

$$1 - \sqrt{\frac{2x}{x^2+1}} \leq \log x^2 - 1 + \frac{1}{x^2}. \tag{29}$$

PROOF OF LEMMA A.1
Let us introduce a new function $f(x)$ as follows,

$$h(x) = \log x^2 + \frac{1}{x^2} + \sqrt{\frac{2x}{x^2+1}}. \tag{30}$$

The claim holds if $h(x) \geq 2$ for all $x > 0$. Note that the first derivative of $h(x)$ could be written as,

$$h'(x) = 2(x^2-1)\left(\frac{1}{x^3} - \frac{1}{2(x^2+1)\sqrt{2x(x^2+1)}}\right).$$

Noting also that,

$$\frac{2(x^2+1)\sqrt{2x(x^2+1)}}{x^3} = 2\sqrt{2}\sqrt{x\left(1+\frac{1}{x^2}\right)}\left(1+\frac{1}{x^2}\right) \geq 2\sqrt{2}\sqrt{x\frac{1}{2x}} \times 1 = 2 > 1.$$

Hence $h'(x) \geq 0$ if $x \geq 1$ and $h'(x) < 0$ otherwise. That is to say, $h(x)$ attains the minimum at $x = 1$. Using the fact that $h(1) = 2$ we then obtain $h(x) \geq 2$ for all $x > 0$. So the proof of this lemma is complete. □

The following lemma states that the order of the Hellinger distance between (β_1, σ_1) and (β_2, σ_2) is controlled by the Euclidean distance of the finite-dimensional parametric parts β_1 and β_2 as well as the uniform norm of the difference on the infinite-dimensional parts σ_1 and σ_2.

LEMMA A.2 *Let $\lambda_{max}(E(X_i X_i'))$ be bounded by some positive constant m_2 for $i = 1, 2, \ldots, n$, then we have,*

$$\begin{aligned} d_H^2(\eta_1, \eta_2) &= 2 - 2\int_{\mathscr{X}} \exp\left\{-\frac{((\beta_1-\beta_2)^T X)^2}{4(\sigma_1^2(x)+\sigma_2^2(x))}\right\} \sqrt{\frac{2\sigma_1(x)\sigma_2(x)}{\sigma_1^2(x)+\sigma_2^2(x)}}\, dG_0(x) \\ &\leq \frac{m_2}{4\underline{\sigma}^2}\|\beta_1-\beta_2\|_2^2 + \frac{1}{4}z\left(\frac{\underline{\sigma}^2}{\overline{\sigma}^2}\right)\frac{\overline{\sigma}^2}{\underline{\sigma}^4}\sup_{x\in\mathscr{X}}|\sigma_1(x)-\sigma_2(x)|^2. \end{aligned} \tag{31}$$

PROOF OF LEMMA A.2

An application of the elementary inequality $1-ab \leq 1-a+1-b$ for $a \leq 1$ and $b \leq 1$ yields,

$$\begin{aligned}
d_H^2(\eta_1,\eta_2) &= 2 - 2\int_{\mathscr{X}} \exp\left\{-\frac{((\beta_1-\beta_2)^T X)^2}{4(\sigma_1^2(x)+\sigma_2^2(x))}\right\}\sqrt{\frac{2\sigma_1(x)\sigma_2(x)}{\sigma_1^2(x)+\sigma_2^2(x)}}\,dG_0(x) \\
&\leq \int_{\mathscr{X}} 2\left(1-\exp\left\{-\frac{((\beta_1-\beta_2)^T X)^2}{4(\sigma_1^2(x)+\sigma_2^2(x))}\right\}\right) + 2\left(1-\sqrt{\frac{2\sigma_1(x)\sigma_2(x)}{\sigma_1^2(x)+\sigma_2^2(x)}}\right) dG_0(x) \\
&\leq \int_{\mathscr{X}} \left\{\frac{((\beta_1-\beta_2)^T X)^2}{2(\sigma_1^2(x)+\sigma_2^2(x))} + \log\left(\frac{\sigma_1^2(x)}{\sigma_2^2(x)}\right) - 1 + \frac{\sigma_2^2(x)}{\sigma_1^2(x)}\right\} dG_0(x) \\
&\leq \frac{1}{4\underline{\sigma}^2}\lambda_{max}(E(X_iX_i'))\|\beta_1-\beta_2\|_2^2 + \frac{1}{4}z\left(\frac{\underline{\sigma}^2}{\overline{\sigma}^2}\right)\frac{\overline{\sigma}^2}{\underline{\sigma}^4}\sup_{x\in\mathscr{X}}|\sigma_1(x)-\sigma_2(x)|^2,
\end{aligned}$$

where the penultimate inequality follows from the elementary inequality $1-e^{-x} \leq x$ for $x \geq 0$ and lemma A.1 in the Appendix. Thus the assertion follows by the assumption $\lambda_{max}(E(X_iX_i')) \leq m_2$ for $i = 1, 2, \ldots, n$. □

The following lemma states that we could bound the first and second moments of log likelihood ratio from above.

LEMMA A.3 *Let $\lambda_{max}(E(X_iX_i')) \leq m_2$, where $m_2 > 0$, then the following inequalities hold.*

$$K(\eta,\eta_0) \leq m_3\left(\sup_{x\in\mathscr{X}}|\sigma(x)-\sigma_0(x)|^2 + \|\beta-\beta_0\|_2^2\right), \tag{32}$$

$$V(\eta,\eta_0) \leq m_4\left(\sup_{x\in\mathscr{X}}|\sigma(x)-\sigma_0(x)|^2 + \|\beta-\beta_0\|_2^2\right). \tag{33}$$

PROOF OF LEMMA A.3

A straightforward computation for $K(\eta,\eta_0)$ shows that,

$$\begin{aligned}
K(\eta,\eta_0) &= \int_{\mathscr{X}}\int_{\mathscr{Y}} f_{x\eta_0}(y)\log\frac{f_{x\eta_0}(y)}{f_{x\eta}(y)}\,dy\,dG_0(x) \\
&= \int_{\mathscr{X}}\int_{\mathscr{Y}} f_{x\eta_0}(y)\frac{1}{2}\left\{\log\frac{\sigma^2(x)}{\sigma_0^2(x)} - \frac{(y-\beta_0'x)^2}{\sigma_0^2(x)} + \frac{(y-\beta'x)^2}{\sigma^2(x)}\right\} dy\,dG_0(x) \\
&= \int_{\mathscr{X}}\frac{1}{2}\left\{\log\frac{\sigma^2(x)}{\sigma_0^2(x)} - 1\right\} dG_0(x) + \int_a^b\int_{\mathscr{Y}} f_{x\eta_0}(y)\left\{\frac{1}{2\sigma^2(x)}(y-\beta_0'x+\beta_0'x-\beta'x)^2\right\} dy\,dG_0(x) \\
&= \int_{\mathscr{X}}\frac{1}{2}\left\{\log\frac{\sigma^2(x)}{\sigma_0^2(x)} - 1 + \frac{\sigma_0^2(x)}{\sigma^2(x)} + \frac{1}{\sigma^2(x)}(\beta_0-\beta)'xx'(\beta_0-\beta)\right\} dG_0(x) \\
&\leq 2z\left(\frac{\underline{\sigma}^2}{\overline{\sigma}^2}\right)\frac{\overline{\sigma}^2}{\underline{\sigma}^4}\sup_{x\in\mathscr{X}}|\sigma(x)-\sigma_0(x)|^2 + \underline{\sigma}^{-2}\lambda_{max}(E(X_iX_i'))\|\beta-\beta_0\|_2^2,
\end{aligned}$$

where the final line follows from lemma A.5. Thus the assertion (32) follows by taking,

$$m_3 = \left\{2z\left(\frac{\underline{\sigma}^2}{\overline{\sigma}^2}\right)\frac{\overline{\sigma}^2}{\underline{\sigma}^4},\ \underline{\sigma}^{-2}\lambda_{max}(E(X_iX_i'))\right\}.$$

For $V(\eta, \eta_0)$, simple algebra delivers that,

$$\begin{aligned} V(\eta, \eta_0) &= \int_{\mathscr{X}} \int_{\mathscr{Y}} f_{x\eta_0}(y) \left(\log \frac{f_{x\eta_0}(y)}{f_{x\eta}(y)} \right)^2 dy \, dG_0(x) \\ &= \int_{\mathscr{X}} \left\{ \left(\frac{\sigma_0^2(x)}{\sigma^2(x)} - 1 \right)^2 + \frac{\sigma_0^4(x)}{\sigma^4(x)} (\beta_0 - \beta)' x x' (\beta_0 - \beta) \right\} dG_0(x) \\ &\leq \underline{\sigma}^{-2} \int_{\mathscr{X}} (\sigma^2(x) - \sigma_0^2(x))^2 \, dG_0(x) + \left(\frac{\overline{\sigma}}{\underline{\sigma}} \right)^4 \lambda_{max}(E(XX')) \|\beta - \beta_0\|_2^2 \\ &\leq \frac{4\overline{\sigma}^2}{\underline{\sigma}^2} \sup_{x \in \mathscr{X}} |\sigma(x) - \sigma_0(x)|^2 + \left(\frac{\overline{\sigma}}{\underline{\sigma}} \right)^4 \lambda_{max}(E(X_i X_i')) \|\beta - \beta_0\|_2^2. \end{aligned}$$

Here we let,

$$m_4 = \left\{ \frac{4\overline{\sigma}^2}{\underline{\sigma}^2}, \left(\frac{\overline{\sigma}}{\underline{\sigma}} \right)^4 \lambda_{max}(E(X_i X_i')) \right\},$$

therefore the assertion (33) follows. □

An immediate consequence from lemma A.3 implies that the following result holds.

COROLLARY A.4 *Under the condition described in lemma A.3, we have,*

$$\max \{K(\eta, \eta_0), V(\eta, \eta_0)\} \leq m_5 \left(\sup_{x \in \mathscr{X}} |\sigma(x) - \sigma_0(x)|^2 + \|\beta - \beta_0\|_2^2 \right), \tag{34}$$

for some positive constant m_5.

LEMMA A.5 *Let* $z(t) = \dfrac{t - 1 - \log t}{(t-1)^2}$ *be a positive decreasing function on* $(0, \infty)$*, then for any* $t \in \left[\frac{\underline{\sigma}^2}{\overline{\sigma}^2}, \frac{\overline{\sigma}^2}{\underline{\sigma}^2} \right]$*, the following inequality holds,*

$$\frac{4\underline{\sigma}^2}{\overline{\sigma}^4} z\left(\frac{\overline{\sigma}^2}{\underline{\sigma}^2} \right) \tilde{d}_2^2(\sigma, \sigma_0) \leq \int_a^b \left(\frac{\sigma_0^2(x)}{\sigma^2(x)} - 1 - \log \frac{\sigma_0^2(x)}{\sigma^2(x)} \right) dG_0(x) \leq \frac{4\overline{\sigma}^2}{\underline{\sigma}^4} z\left(\frac{\underline{\sigma}^2}{\overline{\sigma}^2} \right) \tilde{d}_2^2(\sigma, \sigma_0),$$

where $\tilde{d}_2^2(\sigma, \sigma_0) = \int_a^b (\sigma(x) - \sigma_0(x))^2 \, dG_0(x)$.

PROOF OF LEMMA A.5

Observe that,

$$(t-1)^2 z\left(\frac{\overline{\sigma}^2}{\underline{\sigma}^2} \right) \leq t - 1 - \log t \leq (t-1)^2 z\left(\frac{\underline{\sigma}^2}{\overline{\sigma}^2} \right).$$

Let $t = \frac{\sigma_0^2(X)}{\sigma^2(X)}$ and notice that,

$$\frac{(\sigma^2(X) - \sigma_0^2(X))^2}{\sigma^4(X)} z\left(\frac{\overline{\sigma}^2}{\underline{\sigma}^2} \right) \leq \frac{\sigma_0^2(X)}{\sigma^2(X)} - 1 - \log \frac{\sigma_0^2(X)}{\sigma^2(X)} \leq \frac{(\sigma^2(X) - \sigma_0^2(X))^2}{\sigma^4(X)} z\left(\frac{\underline{\sigma}^2}{\overline{\sigma}^2} \right).$$

Therefore the claim follows by taking expectation with respect to the distribution function $G_0(x)$ on the inequality above. □

LEMMA A.6 *Given* $0 < \alpha \leq q$*, and for each function* $f \in C^\alpha[(0,1)^d]$*, there exists some* $\boldsymbol{\theta} \in \mathbb{R}^{J^d}$ *and a positive constant* C *that depends solely on* q *such that,*

$$\|f - \boldsymbol{\theta}^T \boldsymbol{\xi}\|_\infty \leq C J^{-\alpha} \|D^{(\alpha)} f\|_\infty.$$

Futhermore, if $\underline{\sigma} < f < \overline{\sigma}$*, every element of* $\boldsymbol{\theta}$ *could be chosen to be between* $\underline{\sigma}$ *and* $\overline{\sigma}$.

Proof of lemma A.6

The first part is as same as Lemma 1 of Shen and Ghosal (2012). And the proof of the second part goes throughout the part (b) of Shen and Ghosal (2012) by choosing $\tilde{f} = f - \underline{\sigma}$ and $\tilde{g} = \overline{\sigma} - f$. □

The following two lemmas state that the approximation error of the transform stochastic process could be controlled by the corresponding primitive process with respective to the uniform norm.

Lemma A.7

$$\sup_{x \in \mathscr{X}} \left| \tilde{\Psi}(W(x)) - \tilde{\Psi}(w_0(x)) \right| \leq C \sup_{x \in \mathscr{X}} |W(x) - w_0(x)| . \tag{35}$$

B Proof of Theorems

B.1 Proof of Theorem 4.1

Here we provide the proof of all the results developed in Section 4.

Proof of theorem 4.1

We apply Theorem 4 of Ghosal and van der Vaart (2007a) to prove this theorem in a similar manner as Lin and Dunson (2014). In particular, let,

$$V_n = \{\sigma = \tilde{\Psi}(W) : W \in U_n\}, \tag{36}$$

where U_n is a the measurable subset described in theorem C.5. Now we determine the upper bound on the entropy number on the sieve of the support of the product prior $\Pi_\eta = \Pi_\sigma \times \Pi_\beta$. Define,

$$\mathscr{F}_n = \left\{ (\sigma, \beta) : \sigma \in V_n,\, \beta \in [\underline{\beta}, \overline{\beta}]^d \right\}. \tag{37}$$

Since $\tilde{\Psi}$ is a one to one map from $\mathbb{R}$ to $[\underline{\sigma}, \overline{\sigma}]$, then $V_n \subset B_n$. Hence the number of $\bar{\varepsilon}_n$-balls needed to cover V_n is less than B_n in terms of the uniform distance. That is,

$$\log N(\varepsilon_n, V_n, \|\cdot\|_\infty) \leq \log N(\varepsilon_n, B_n, \|\cdot\|_\infty), \tag{38}$$

which is bounded by $Dn\varepsilon_n^2$ by (73). To bound from above the entropy number on $\mathscr{F}_n$, we consider the covering number of the one dimensional set $\{\beta_1 : \beta_1 \in [\underline{\beta}, \overline{\beta}]\}$. Let $N = \left\{ \left[\frac{\overline{\beta} - \underline{\beta}}{2\varepsilon_n} \right] + 1 \right\}$, the interval $[\underline{\beta}, \overline{\beta}]$ could be partitioned into N sub-intervals with the equal length $\frac{\overline{\beta} - \underline{\beta}}{N}$. We denote all the middle points of these equidistant intervals by the set,

$$T = \left\{ \underline{\beta} + i \frac{\overline{\beta} - \underline{\beta}}{2N} : i = 1, 3, \ldots, 2N - 1 \right\}.$$

Then every equidistant interval could be covered by one neighborhood of some point in T with radius $\bar{\varepsilon}_n$. Thus the covering number of the set $\{\beta : \beta \in [\underline{\beta}, \overline{\beta}]^d\}$ is,

$$N\left(\varepsilon_n d^{1/2},\, [\underline{\beta}, \overline{\beta}]^d,\, \|\cdot\|_2 \right) \leq N^d.$$

In view of (31), observe that if $\sup_{x\in\mathscr{X}} |\sigma(x) - \sigma_0(x)| \leq C\varepsilon_n$ and $\|\beta - \beta_0\|_2 \leq \varepsilon_n d^{1/2}$, then we have that,

$$\begin{aligned} d_H^2(\eta, \eta_0) &\lesssim \sup_{x\in\mathscr{X}} |\sigma(x) - \sigma_0(x)|^2 + \|\beta - \beta_0\|_2^2, \\ &\leq \varepsilon_n (C^2 + d)^{1/2}. \end{aligned}$$

Therefore, the $\varepsilon_n(C^2+d)^{1/2}$-covering number of $\mathscr{F}_n$ is bounded by $e^{Dn\varepsilon_n^2} \times N^d$, that is,

$$\log N\left(\varepsilon_n (C^2+d)^{1/2},\, \mathscr{F}_n,\, d_H\right) \leq Dn\varepsilon_n^2 + \log N.$$

Using the assumption $\log\left\{\left[\frac{\bar{\beta}-\underline{\beta}}{2\varepsilon_n}\right] + 1\right\} \leq n\varepsilon_n^2$ we obtain,

$$\log N\left((C^2+d)^{1/2}\varepsilon_n, \mathscr{F}_n, d_H\right) \leq (D+d)n\varepsilon_n^2.$$

We proceed to show that the prior Π_η assigns a large amount of probability mass on some specialized Kullback-Leibler ball of the true value η_0. Let,

$$B^*(\eta_0, \varepsilon_n) = \{\eta :\, K(\eta, \eta_0) < \varepsilon_n^2, V(\eta, \eta_0) < \varepsilon_n^2\}. \tag{39}$$

We need to bound from below $\Pi(B^*(\eta_0, \varepsilon_n))$. By corollary A.4, it follows that,

$$B^*(\eta_0, \varepsilon_n) \supset \left\{\eta = (\beta, \sigma) :\, \|\beta - \beta_0\|_2 \leq \frac{\tilde{D}\varepsilon_n}{2},\, \|\sigma - \sigma_0\|_\infty \leq \frac{\tilde{D}\varepsilon_n}{2}\right\}. \tag{40}$$

for some constant $\tilde{D}$. Therefore the prior mass on $B^*(\eta_0, \varepsilon_n)$ could be lower bounded by,

$$\Pi_\sigma\left(\|\sigma - \sigma_0\|_\infty \leq \frac{\tilde{D}\varepsilon_n}{2}\right) \times \Pi_\beta\left(\|\beta - \beta_0\|_2 \leq \frac{\tilde{D}\varepsilon_n}{2}\right).$$

Applying lemma A.7 gives rise to,

$$\Pi_\sigma\left(\|\sigma - \sigma_0\|_\infty \leq \frac{\tilde{D}\varepsilon_n}{2}\right) \geq \Pi_W\left(\|W - w_0\|_\infty \leq \frac{\tilde{D}\varepsilon_n}{2C}\right),$$

which is greater than $\exp\left\{-\frac{\tilde{D}^2 n\varepsilon_n^2}{16C^2}\right\}$. In view of the assumption on the prior of β, we have ,

$$\begin{aligned} \Pi(B^*(\eta_0, \varepsilon_n)) &\geq \Pi_\sigma\left(\|\sigma - \sigma_0\|_\infty \leq \frac{\tilde{D}\varepsilon_n}{2}\right) \times \Pi_\beta\left(\|\beta - \beta_0\|_2 \leq \frac{\tilde{D}\varepsilon_n}{2}\right), \\ &\geq \exp\left\{-\frac{\tilde{D}^2 n\varepsilon_n^2}{16C^2}\right\} \times \exp(-\bar{D}n\varepsilon_n^2), \\ &\geq \exp(-\tilde{D}_1 n\varepsilon_n^2), \end{aligned}$$

for some positive constant $\tilde{D}_1$.

It remains to show that prior on the complement of the sieve is negligible. In fact, since $\{\eta :\, \eta \notin \mathscr{F}_n\} \subset \{\eta :\, \sigma \notin V_n\}$, it is easy to say, by (72),

$$\Pi_\eta\{\eta :\, \eta \notin \mathscr{F}_n\} \subset \Pi_\sigma\{\eta :\, \sigma \notin V_n\} \subset \Pi_W\{W :\, W \notin U_n\} \leq \exp\{-n\varepsilon_n^2\}. \tag{41}$$

So the claim follows since all the three key conditions listed in Theorem 4 of Ghosal and van der Vaart (2007a) are satisfied. □

In order to prove theorem 4.2, we first present a variant of main results stated in Shen and Ghosal (2012) in the following two technical lemmas.

LEMMA B.1 *Let,*

$$\tilde{V}_{J_n,M_n} = \{\sigma = \tilde{\Phi}(W^{J,\boldsymbol{\theta}}) : W^{J,\boldsymbol{\theta}} = \boldsymbol{\theta}^T\boldsymbol{\xi},\, \boldsymbol{\theta} \in \mathbb{R}^j,\, j \le J_n,\, \|\boldsymbol{\theta}\|_\infty \le M_n\}, \tag{42}$$

$$\tilde{\mathcal{W}}_{J_n,M_n} = \{(\sigma,\beta) : \sigma \in \tilde{V}_{J_n,M_n},\, \beta \in [\underline{\beta},\overline{\beta}]^d\}, \tag{43}$$

$$d_2(\eta,\eta_0) = \left\{\int_0^1 [\sigma(x)-\sigma_0(x)]^2\, dG_0(x)\right\}^{1/2} + \|\beta-\beta_0\|_2. \tag{44}$$

Assume that the conditions listed in Theorem 1 *of Shen and Ghosal (2012) hold relative to uniform metric* $\|\cdot\|_\infty$*, then for some positive constants* $\tilde{a}_1$, $\tilde{a}_2$, $\tilde{b}$*, we have the following,*

$$\log D(\varepsilon_n, \tilde{\mathcal{W}}_{J_n,M_n}, d_2) \le n\varepsilon_n^2, \tag{45}$$

$$\Pi(W \notin \tilde{\mathcal{W}}_{J_n,M_n}) \le \tilde{a}_1 \exp\{-\tilde{b} n\bar{\varepsilon}_n^2\}, \tag{46}$$

$$-\log \Pi\{\eta = (\sigma,\beta) : \|\sigma-\sigma_0\|_\infty^2 + \|\beta-\beta_0\|_2^2 \le \bar{\varepsilon}_n^2\} \le \tilde{a}_2 n\bar{\varepsilon}_n^2. \tag{47}$$

PROOF OF LEMMA B.1

We omit the proof of assertions (45) and (46) since it is similar to the corresponding parts in the proof of theorem 4.1. We are in a position to show (47). Observe that,

$$\begin{aligned}
&\Pi\{\eta = (\sigma,\beta) : \|\sigma-\sigma_0\|_\infty^2 + \|\beta-\beta_0\|_2^2 \le \bar{\varepsilon}_n^2\},\\
&\quad \ge \Pi_\sigma\left(\|\sigma-\sigma_0\|_\infty \le \frac{\bar{\varepsilon}_n}{2}\right) \times \Pi_\beta\left(\|\beta-\beta_0\|_2 \le \frac{\bar{\varepsilon}_n}{2}\right),\\
&\quad \ge \Pi_w\left(\|w-w_0\|_\infty \le \frac{\bar{\varepsilon}_n}{2}\right) \times \exp(-cd\log(1/\bar{\varepsilon}_n)),\\
&\quad \ge \exp\left\{-a_2 n\bar{\varepsilon}_n^2\right\} \times \exp(-\tilde{b}_2 n\bar{\varepsilon}_n^2),\\
&\quad \ge \exp(-\tilde{a}_2 n\bar{\varepsilon}_n^2),
\end{aligned}$$

where $\tilde{a}_2 = a_2 + \tilde{b}_2$. The assertion (47) follows by taking logarithm transformation on both sides above. We thus complete the proof of this lemma. □

LEMMA B.2 *Suppose that the conditions except (68) listed in Theorem 2 of Shen and Ghosal (2012) hold for the case* $r = \infty$*, then the posterior distribution of* η *converges at rate* ε_n *with respective to the Hellinger distance.*

PROOF OF LEMMA B.2

Notice that $K(p_{f_0}, p_f)$ and $V(p_{f_0}, p_f)$ exhibited in Theorem 2 of Shen and Ghosal (2012) are essentially the same as $K(\eta_0,\eta)$ and $V(\eta_0,\eta)$ respectively described in (1) and (2). We employ the similar arguments in the proof of Theorem 2 in Shen and Ghosal (2012) to show this lemma. It suffices to show that the following conditions stated in Theorem 4 of Ghosal

and van der Vaart (2007a).

$$\log D(\varepsilon_n, \tilde{\mathcal{W}}_{J_n,M_n}, d_H) \leq b_1 n\varepsilon_n^2, \tag{48}$$

$$\Pi(W \notin \tilde{\mathcal{W}}_{J_n,M_n}) \leq b_3 \exp\{-(b_2+4)n\bar{\varepsilon}_n^2\}, \tag{49}$$

$$\Pi(B^*(\eta_0, \bar{\varepsilon}_n)) \geq b_4 \exp\{-b_2 n\bar{\varepsilon}_n^2\}, \tag{50}$$

for some positive constants b_1, b_2, b_3, b_4, where $\tilde{\mathcal{W}}_{J_n,M_n}$ is described in lemma B.1 and $B^*(\eta_0, \bar{\varepsilon}_n) = \{\eta : K(\eta, \eta_0) < \varepsilon_n^2, V(\eta, \eta_0) < \bar{\varepsilon}_n^2\}$. It is easy to show (48) and (49) by the same arguments used in the proof of Theorem 2 in Shen and Ghosal (2012). Now it remains to check (50). In fact, observe that by corollary A.4,

$$B^*(\eta_0, \bar{\varepsilon}_n) \supset \{\eta = (\sigma, \beta) : \|\sigma - \sigma_0\|_\infty^2 + \|\beta - \beta_0\|_2^2 \leq \bar{\varepsilon}_n^2\}.$$

It follows that by (47) in lemma B.1,

$$\begin{aligned}\Pi(B^*(\eta_0, \bar{\varepsilon}_n)) &\geq \Pi\{\eta = (\sigma, \beta) : \|\sigma - \sigma_0\|_\infty^2 + \|\beta - \beta_0\|_2^2 \leq \bar{\varepsilon}_n^2\} \\ &\geq \exp(-\tilde{a}_2 n\bar{\varepsilon}_n^2).\end{aligned}$$

Then the proof of this lemma is complete. □

B.2 Proof of Theorem 4.2

Proof of theorem 4.2

In order to obtain the rate ε_n like this, we only need to apply lemma B.2 with the appropriate choice of $\bar{J}_n$, J_n, M_n, $\bar{\varepsilon}_n$. It is easy to say that (66) and (67) described in Theorem 2 of Shen and Ghosal (2012) in terms of tensor-product spline basis. An application of corollary A.4 yields that,

$$\max(K(\eta_0, \eta),\, V(\eta_0, \eta)) \preceq (\|\sigma - \sigma_0\|_\infty^2 + \|\beta - \beta_0\|_2^2).$$

Meanwhile, lemma 3.1 implies the approximation error $e(J) \approx J^{-\alpha}$. We proceed to determine the rate ε_n as follows. Firstly, it follows that $\bar{J}_n^{-\alpha} \leq \bar{\varepsilon}_n$ and $\bar{J}_n \log n \leq n\bar{\varepsilon}_n^2$ by (67). Hence we can choose $M_n = n^{1/t_3}$, $\bar{J}_n = (n/\log n)^{1/(2\alpha+d)}$ and $\bar{\varepsilon}_n = (n/\log n)^{-\alpha/(2\alpha+d)}$. Observe that $n\bar{\varepsilon}_n^2 \preceq J_n \log^{t_1} n$ by (65), we can also choose $J_n = n^{1/(2\alpha+d)}(\log n)^{2\alpha/(2\alpha+d)-t_2}$. Noting also that $J_n \log n \preceq n\varepsilon_n^2$ by (66), so that we get the rate ε_n as $n^{-\alpha/(2\alpha+d)}(\log n)^{\alpha/(2\alpha+d)-(t_2-1)/2}$. Then the proof of this theorem is complete. □

C Auxiliary theorems for this paper

For easy reference, we collect some complementary results in the literature in aid of the proof of the theorems in this present article.

Theorem C.1 (Ghosal and van de Vaart (2007)) *Let $P_\theta^{(n)}$ be product measures and d_n be defined as follows:*

$$d_n(\theta, \theta') = \frac{1}{n}\int (\sqrt{p_{\theta,i}} - \sqrt{p_{\theta',i}})^2 \, d\mu_i. \tag{51}$$

Suppose that for a sequence $\varepsilon_n \to 0$ such that $n\varepsilon_n^2$ is bounded away from zero, some $k > 1$, all sufficiently large j and sets $\Theta_n \subset \Theta$, the following conditions hold:

$$\sup_{\varepsilon > \varepsilon_n} \log N(\varepsilon/36, \{\theta \in \Theta_n : d_n(\theta, \theta_0) < \varepsilon\}, d_n) \leq n\varepsilon_n^2, \tag{52}$$

$$\frac{\Pi_n(\Theta \backslash \Theta_n)}{\Pi_n(B_n^*(\theta_0, \varepsilon_n; k))} = o(e^{-2n\varepsilon_n^2}), \tag{53}$$

$$\frac{\Pi_n(\theta \in \Theta_n : \ j\varepsilon_n \leq d_n(\theta, \theta_0) \leq 2j\varepsilon_n)}{\Pi_n(B_n^*(\theta_0, \varepsilon_n; k))} \leq e^{n\varepsilon_n^2 j^2/4}. \tag{54}$$

Then $P_\theta^{(n)} \Pi_n(\theta : d_n(\theta, \theta_0) \geq M_n \varepsilon_n | X^{(n)}) \to 0$ for every $M_n \to \infty$.

Lemma C.2 (Shen and Ghosal (2012)) *For any $1 \leq p \leq \infty$, we have,*

$$\|\boldsymbol{\theta}_1^T \boldsymbol{\xi} - \boldsymbol{\theta}_2^T \boldsymbol{\xi}\|_r \leq \sum_{j=1}^{J} |\theta_{1j} - \theta_{2j}| \max_{1 \leq j \leq J} \|\xi_j\|_p \leq \sqrt{J} \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2 \, C_{p,J}, \tag{55}$$

where,

$$C_{p,J} \equiv \max_{1 \leq j \leq J} \|\xi_j\|_p \asymp \begin{cases} 1 & p = 2 \\ \sqrt{J} & p = \infty \end{cases}$$

Theorem C.3 (Shen and Ghosal (2012)) *Let $\varepsilon_n \geq \bar{\varepsilon}_n$ be two sequence of positive numbers satisfying $\varepsilon_n \to 0$ and $n\bar{\varepsilon}_n^2 \to \infty$ as $n \to \infty$. For a function w_0, suppose that there exist sequences of positive numbers J_n, $\bar{J}_n$ and M_n, a strictly decreasing, nonnegative function $e(\cdot)$ and a $\theta_{0,j} \in \mathbb{R}^j$ for any $j \in \mathbb{N}$, such that the following conditions hold for some positive constants a_1, a_1', a_2:*

$$\|\theta_{0,j}\| \leq H, \ d_2(w_0, \boldsymbol{\theta}_{0,j}^T \boldsymbol{\xi}) \leq e(j), \tag{56}$$

$$J_n\{\log J_n + \log a(J_n) + \log M_n + \log(1/\varepsilon_n)\} \leq n\varepsilon_n^2, \tag{57}$$

$$e(\bar{J}_n) \leq \bar{\varepsilon}_n, \ \log\{1/B(\bar{J}_n)\} + c_2 \bar{J}_n \log(2b(\bar{J}_n)/\bar{\varepsilon}_n) \leq a_2 n\bar{\varepsilon}_n^2, \tag{58}$$

$$A(J_n) \leq a_1 \exp\{-(a_2+4)n\bar{\varepsilon}_n^2\}, \ J_n \exp\{-CM_n^{t_3}\} \leq a_1' \exp\{-(a_2+4)n\bar{\varepsilon}_n^2\}. \tag{59}$$

Let $\mathcal{W} = \{w = \boldsymbol{\theta}^T \boldsymbol{\xi} : \ \boldsymbol{\theta} \in \mathbb{R}^j, j \leq J_n, \|\boldsymbol{\theta}\|_\infty \leq M_n\}$. Then the following assertions hold:

$$\log D(\varepsilon_n, \mathcal{W}_{J_n, M_n}, d_2) \leq n\varepsilon_n^2, \tag{60}$$

$$\Pi(W \notin \mathcal{W}_{J_n, M_n}) \leq (a_1 + a_1') \exp\{-(a_2+4)n\bar{\varepsilon}_n^2\}, \tag{61}$$

$$-\log \Pi\{w = \boldsymbol{\theta}^T \boldsymbol{\xi} : \ d_2(w_0, w) \leq \bar{\varepsilon}_n\} \leq a_2 n\bar{\varepsilon}_n^2. \tag{62}$$

Theorem C.4 (Shen and Ghosal (2012)) *Suppose that we have independent observations X_i following some distributions with densities $p_{i,w} : i = 1, \ldots, n$ respectively. Let $w_0 \in C^\alpha(\Omega_0)$ be the true value of w. let r be either 2 or ∞. Let $\varepsilon_n \geq \bar{\varepsilon}_n$ be two sequences of positive numbers satisfying $\varepsilon_n \to 0$ and $n\bar{\varepsilon}_n^2 \to \infty$ as $n \to \infty$. Assume that there exists a $\boldsymbol{\theta}_0 \in \mathbb{R}^J$, $\|\boldsymbol{\theta}_0\| \leq H$ and some positive constants C_1, C_2 satisfying,*

$$\|w_0 - \boldsymbol{\theta}_0^T \boldsymbol{\xi}\|_r \leq C_1 J^\alpha (\log J)^s, \ s \geq 0, \tag{63}$$

$$\|\boldsymbol{\theta}_1^T \boldsymbol{\xi} - \boldsymbol{\theta}_2^T \boldsymbol{\xi}\|_r \leq C_2 J^{K_0} \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2, \ K_0 \geq 0, \ \text{for any } \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^J. \tag{64}$$

Assume that the prior on J and $\boldsymbol{\theta}$ satisfy some conditions $(A2)$ and $(A3)$ in their paper. Let J_n, $\bar{J}_n \geq 2$ and M_n be sequences of positive numbers such that the following hold:

$$J_n \log^{t_1} J_n \geq 6n\bar{\varepsilon}_n^2,\ \log J_n + 6n\varepsilon_n^2 \leq c_1 M_n^{t_3}, \tag{65}$$

$$J_n\{(K_0+1)\log J_n + \log M_n + \log(1/\varepsilon_n) + \log n\} \leq n\varepsilon_n^2, \tag{66}$$

$$\bar{J}_n^{-\alpha}(\log \bar{J}_n)^s \leq \bar{\varepsilon}_n,\ \bar{J}_n\{\log^{t_2} \bar{J}_n + c_2 K_0 \log(\bar{J}_n) + c_2 \log(1/\bar{\varepsilon}_n)\} \leq 2n\bar{\varepsilon}_n^2, \tag{67}$$

$$\rho_n(w_1, w_2) \lesssim n^{C_3}\|w_1 - w_2\|_r \text{ for any } w_1,\ w_2 \in \mathcal{W}_{J_n, M_n} \text{ and some constant } C_3 > 0, \tag{68}$$

$$\max_{1\leq i\leq n}\{\, K(p_{i,w_0}, p_{i,w}),\ V(p_{i,w_0}, p_{i,w})\,\} \lesssim \|w_1 - w_2\|_r, \tag{69}$$

provided $\|w_1 - w_2\|_r$ is sufficiently small. Then the posterior of w converges around w_0 at the rate ε_n with respect to ρ_n.

THEOREM C.5 (de Jonge and van Zanten (2012)) *Suppose that for every $m \geq 1$,*

$$C_1 \exp(-D_1 m^d \log^t m) \leq P(M = m) \leq C_2 \exp(-D_2 m^d \log^t m), \tag{70}$$

for some constants $C_1, C_2, D_1, D_2, t \geq 0$. If $w_0 \in C^r([0,1]^d)$ for some integer $r \leq q$, then there exists for every constant $C > 0$, a constant $D > 0$ and measurable subsets U_n of $C([0,1]^d)$ such that,

$$P\left(\|W - w_0\|_\infty \leq 2\varepsilon_n\right) \geq \exp\left(-n\varepsilon_n^2\right), \tag{71}$$

$$P(W \notin U_n) \leq \exp\left(-Cn\varepsilon_n^2\right), \tag{72}$$

$$\log N(2\bar{\varepsilon}_n, U_n, \|\cdot\|_\infty) \leq Dn\bar{\varepsilon}_n^2, \tag{73}$$

are satisfied for sufficiently large n, and for ε_n and $\bar{\varepsilon}_n$ given by,

$$\varepsilon_n = c(n/\log^{1\vee t} n)^{-\frac{r}{d+2r}} \qquad \bar{\varepsilon}_n = n^{-\frac{r}{d+2r}}(\log n)^{\frac{(1\vee t)r}{d+2r} + (\frac{1-t}{2})+}, \tag{74}$$

for $c > 0$ a large enough constant.

References

Belitser, E. and S. Ghosal (2003). Adaptive Bayesian inference on the mean of an infinite-dimensional normal distribution. *The Annals of Statistics 31*(2), 536–559.

Belitser, E. and P. Serra (2014). Adaptive priors based on splines with random knots. *Bayesian Analysis 9*(4), 859–882.

Chib, S. and E. Greenberg (2013). On conditional variance estimation in nonparametric regression. *Statistics and Computing 23*(2), 261–270.

de Boor, C. (2001). *A Practical Guide to Splines*. New York: Springer.

de Jonge, R. and J. H. van Zanten (2010). Adaptive nonparametric Bayesian inference using location-scale mixture priors. *The Annals of Statistics 38*(6), 3300–3320.

de Jonge, R. and J. H. van Zanten (2012). Adaptive estimation of multivariate functions using conditionally Gaussian tensor-product spline priors. *Electronic Journal of Statistics 6*, 1984–2001.

Ghosal, S., J. Ghosh, and A. W. van der Vaart (2000). Convergence rates of posterior distributions. *The Annals of Statistics 28*(2), 500–531.

Ghosal, S., J. K. Ghosh, and R. Ramamoorthi (1999). Posterior consistency of Dirichlet mixtures in density estimation. *The Annals of Statistics 27*(1), 143–158.

Ghosal, S. and A. W. van der Vaart (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *The Annals of Statistics 29*(5), 1233–1263.

Ghosal, S. and A. W. van der Vaart (2007a). Convergence rates of posterior distributions for noniid observations. *The Annals of Statistics 35*(1), 192–223.

Ghosal, S. and A. W. van der Vaart (2007b). Posterior convergence rates of Dirichlet mixtures at smooth densities. *The Annals of Statistics 35*(2), 697–723.

Goldberg, P. W., C. K. Williams, and C. M. Bishop (1997). Regression with input-dependent noise: A Gaussian process treatment. *Advances in Neural Information Processing Systems 10*, 493–499.

Huang, T.-M. (2004). Convergence rates for posterior distributions and adaptive estimation. *The Annals of Statistics 32*(4), 1556–1593.

Kruijer, W., J. Rousseau, and A. W. van der Vaart (2010). Adaptive Bayesian density estimation with location-scale mixtures. *Electronic Journal of Statistics 4*, 1225–1257.

Lin, L. and D. B. Dunson (2014). Bayesian monotone regression using Gaussian process projection. *Biometrika 101*(2), 303–317.

Norets, A. (2015). Bayesian regression with nonparametric heteroskedasticity. *Journal of Econometrics 185*(2), 409–419.

Norets, A. and D. Pati (2014). Adaptive Bayesian estimation of conditional densities. *arXiv preprint arXiv:1408.5355*.

Pelenis, J. (2014). Bayesian regression with heteroscedastic error density and parametric mean function. *Journal of Econometrics 178*(3), 624–638.

Rousseau, J. (2010). Rates of convergence for the posterior distributions of mixtures of betas and adaptive nonparametric estimation of the density. *The Annals of Statistics 38*(1), 146–180.

Schumaker, L. (2007). *Spline Functions: Basic Theory*. New York: Cambridge University Press.

Scricciolo, C. (2006). Convergence rates for Bayesian density estimation of infinite-dimensional exponential families. *The Annals of Statistics 34*(6), 2897–2920.

Shen, W. and S. Ghosal (2012). MCMC-free adaptive Bayesian procedures using random series prior. *arXiv preprint arXiv:1204.4238*.

Shen, W. and S. Ghosal (2014). Adaptive Bayesian density regression for high-dimensional data. *arXiv preprint arXiv:1403.2695*.

Shen, W., S. T. Tokdar, and S. Ghosal (2013). Adaptive Bayesian multivariate density estimation with Dirichlet mixtures. *Biometrika 100*(3), 623–640.

Stone, C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation. *The Annals of Statistics 22*(1), 118–171.

van der Vaart, A. W. and J. H. van Zanten (2009). Adaptive Bayesian estimation using a Gaussian random field with inverse Gamma bandwidth. *The Annals of Statistics 37*(5), 2655–2675.

Wang, L. (2013). Consistency of posterior distributions for heteroscedastic nonparametric regression models. *Communications in Statistics-Theory and Methods 42*(15), 2731–2740.

Yau, P. and R. Kohn (2003). Estimation and variable selection in nonparametric heteroscedastic regression. *Statistics and Computing 13*(3), 191–208.